Quantifying Paradigm Shape in Spanish Verbs

Undergraduate Honors Research Thesis

Presented in Partial Fulfillment of the Requirements for graduation "with Honors Research Distinction in Linguistics" in the undergraduate colleges of The Ohio State University

by

Grace LeFevre

The Ohio State University

April 2021

Project Advisors:

Professor Andrea Sims, Department of Linguistics, Department of Slavic and East European Languages and Cultures

Professor Micha Elsner, Department of Linguistics

Abstract

This thesis computationally models "paradigm shape," a type of morphological structure that I define by the implicative relations holding among the forms in an inflectional system. Since implicative structure binds the forms in an inflectional system together (Wurzel, 1989), paradigm shape reflects the predictable ways that allomorphs occur in parallel paradigm cells across inflection classes in some languages. Maiden (2005)'s analysis of how certain Romance verbs changed over time in order to conform to existing paradigm shapes highlights the significance of this structure as a historical and cognitive organizing principle. However, paradigm shape has not been computationally formalized in a gradient or replicable way.

Using information-theoretic entropy as defined by Shannon (1948), I develop a method to quantify paradigm shape and I apply it to Spanish verbs as a test case. The method bridges the gap between formal work on the organization of the stem space (e.g. Maiden, 2005; Boyé and Cabredo Hofherr, 2006) and computational work on quantifying predictability in inflectional systems (e.g. Ackerman and Malouf, 2013; Stump and Finkel, 2015). In doing so, it jointly models the distributions of stems and affixes to compute sets of values that characterize the shapes of Spanish verb classes. Comparison of these values across classes captures partial parallelism between them, enabling identification of both allomorphic and distributional class structures (Baerman et al., 2017). These results with Spanish verbs highlight that my method provides a computational means of capturing multiple aspects of inflection class structure in a way that is replicable and extendable to other languages. Potential directions for future work include testing the limits of the method's usefulness on known morphologically difficult systems and applying the method to other Romance languages at various stages of historical development.

Table of Contents

1 Introduction **3**

2 Defining Paradigm Shape **4**

2.1 Paradigm Shape in Spanish Verbs 5

2.2 Previous Work 6

2.2.1 Stem Space Organization 7

2.2.2 Inflectional Complexity 9

3 Method to Quantify Paradigm Shape **11**

3.1 Method Overview 12

3.1.1 Segmentation 12

3.1.2 Maximal Confusion Sets 14

3.1.3 Entropy 15

3.2 Details of Algorithm 16

3.3 Toy Example 18

3.4 Deidentification 20

4 Results with Spanish Verbs **22**

5 Conclusion **26**

Acknowledgements **27**

References **28**

Appendix A: Dendrograms **31**

1 Introduction

This research is grounded in the view of inflectional paradigms as networks of implicative relations (Wurzel, 1989; Bonami and Beniamine, 2016; Ackerman and Malouf, 2016). Wurzel (1989) identified this implicative structure based on the observation that "there are no paradigms (except highly extreme cases of suppletion) that are not based on implications valid beyond the individual word" (p. 114). Implicative relations therefore bind the forms in a paradigm together. For example, taking the small subset of Spanish verb data in Table 1, I can identify the implicative relation "If the PRS.2SG is X*as*, then the PRS.3SG is X*a*." This implication enables the prediction of an unobserved form of a lexeme from an observed form. If I have observed that the PRS.2SG of CANTAR is *cantas*, this implicative relation licenses a prediction that the PRS.3SG of CANTAR is *canta*. This is a simple example applied to only a small number of forms, but the same principles apply to inflection systems as a whole.

In this thesis, I am interested in modeling "paradigm shape" in Spanish verbs, which I define as the network of implicative relations holding among the forms in an inflectional system. As I illustrated with a simple example, this morphological structure reflects the predictable, patterned ways in which stem[1] alternants (and even fully suppletive allomorphs) occur in parallel paradigm cells across inflection classes in languages like Spanish. Historically, some Romance verbs shifted to better conform to existing paradigm shapes, indicating that this is an organizing principle with important implications for language learning and change (Maiden, 2005). However, paradigm shape has not yet been formally quantified in a gradient manner.

I develop a computational method[2] to precisely quantify similarity in paradigm shape and implement it on Spanish verbs. Building on previous work using information theory to quantify predictability (e.g. Ackerman and Malouf, 2013), I apply entropy as defined by Shannon (1948) to Spanish verb forms and compute sets of values characterizing the shapes of the inflection classes.

[1] In Spanish linguistics, the term "stem" is generally used to mean the the root plus the theme vowel. Because of the segmentation strategy I use in this research, what I call the "stem" does not necessarily include the theme vowel and my usage of the term throughout reflects this.

[2] The code and data for this project are available at `github.com/gracelefevre/paradigm-shape`

LEXEME	GLOSS	PRS.1SG	PRS.2SG	PRS.3SG	PRS.1PL	PRS.2PL	PRS.3PL
CANTAR	'sing'	canto	cantas	canta	cantamos	cantáis	cantan
SUBIR	'rise'	subo	subes	sube	subimos	subís	suben
PENSAR	'think'	pienso	piensas	piensa	pensamos	pensáis	piensan
MOVER	'move'	muevo	mueves	mueve	movemos	movéis	mueven
SENTIR	'feel'	siento	sientes	siente	sentimos	sentís	sienten

Table 1: Five Spanish microclasses from my data, shown with their present indicative forms and highlighted to indicate cells with stem alternations.

Substantial previous work has focused exclusively on stem space organization, modeling the distribution of stem alternants within a paradigm (Maiden, 2005; Boyé and Cabredo Hofherr, 2006). My method expands on this work by focusing equally on stems and affixes. Furthermore, my results enable direct analysis of both allomorphic and distributional classes (Baerman et al., 2017), where most previous shape-based analysis has been purely distributional. As such, this work provides a unified, computational approach to phenomena relating to paradigm shape that have predominantly been treated separately in the past.

My thesis is divided into five sections. In section 2, I explain the concept of paradigm shape in Spanish verbs and give an overview of the previous work this research is based on. In section 3, I detail the computational method I developed and how it was implemented in Python. In section 4, I discuss the results the method generated for Spanish verbs. In section 5, I describe the significance of these results and potential directions for future research.

2 Defining Paradigm Shape

In this section, I explain the phenomenon of paradigm shape that this research models. Section 2.1 elaborates on the notion of paradigm shape in Spanish verbs and section 2.2 details previous work that this research builds upon.

2.1 Paradigm Shape in Spanish Verbs

Table 1 shows the present indicative forms of five Spanish verb microclasses from my data. Spanish verbs are traditionally categorized into three classes based on the theme vowel that appears in their infinitive: *-ar* verbs, *-er* verbs, and *-ir* verbs (Butt et al., 2019). In Table 1, CANTAR 'sing' and PENSAR 'think' fall in the *-ar* category, MOVER 'move' in the *-er* category, and SUBIR 'rise' and SENTIR 'feel' in the *-ir* category. In the terminology of Beniamine et al. (2017), these categories are inflectional macroclasses. That is, the lexemes they group together are similar but do not have precisely identical inflectional exponence. In fact, no verb in Table 1 has exactly the same exponence as any other, which means that each one represents a distinct inflectional microclass (again in the sense of Beniamine et al., 2017).

Despite these microclasses all being distinct from one another, similarities clearly exist between them that are not captured by the traditional theme vowel distinctions. For example, PENSAR, MOVER, and SENTIR all have stem alternations, while CANTAR and SUBIR do not. Furthermore, all three lexemes have the same distribution of stem alternants, which are highlighted by the blue cells in Table 1. Even though PENSAR and SENTIR have an *e∼ie* alternation and MOVER has an *o∼ue* alternation, the alternants occur in the same cells across all three verbs. From this perspective, the stem alternations define a different set of inflectional macroclasses for this small subset of Spanish verb data: one that includes CANTAR and SUBIR, and another that includes PENSAR, MOVER, and SENTIR.

The particular alternation present in these verbs follows one of the stem alternant distributions in Romance verbs identified by Maiden (2005). The distribution shown in Table 1, where PENSAR, MOVER, and SENTIR all share an alternation in the 1SG, 2SG, 3SG, and 3PL of the present indicative, is what Maiden termed the "N-pattern." The other patterns he identified include the "L-pattern" (a shared alternation in the 1SG present indicative and all present subjunctive forms) and the "U-pattern" (a shared alternation in the 1SG and 3PL present indicative and all present subjunctive forms).[3] An example of the L-pattern is shown in Table 2. Throughout the history of

[3]The U-pattern is not found in modern Spanish, as most Romance varieties replaced the U-pattern with the L-

PRS.IND.1SG	PRS.IND.2SG	PRS.IND.3SG	PRS.IND.1PL	PRS.IND.2PL	PRS.IND.3PL
valgo	vales	vale	valemos	valéis	valen
PRS.SBJV.1SG	PRS.SBJV.2SG	PRS.SBJV.3SG	PRS.SBJV.1PL	PRS.SBJV.2PL	PRS.SBJV.3PL
valga	valgas	valga	valgamos	valgáis	valgan

Table 2: Present indicative and present subjunctive forms of VALER, 'cost,' highlighted to indicate Maiden (2005)'s L-pattern.

Romance, these distributional patterns have been maintained, reinforced, and applied to new verbs; in doing this, Romance speakers have "pass[ed] up golden opportunities to align allomorphs with morphosyntactic properties" (Maiden, 2005, p.169). For example, most Romance varieties display an N-shaped distribution in suppletively conflating two or three different lexemes for 'go' (Maiden, 2005, p.153), as seen in the Old Spanish present indicative forms of IRE 'go': *vo, vas, va, imos, ides, van*. Though the alternation has been leveled in Modern Spanish, this example shows that paradigmatic distributions of this sort are a cognitively real organizing principle for speakers.

For the five Spanish verbs in Table 1, then, at least two inflectional macroclass groupings are possible: one defined by the theme vowels, and one by the stem alternants. For a given verb, both of these underlying structures impact how predictable other inflected forms of the same lexeme are. In this thesis, I use the term "paradigm shape" to encompass both stem alternations and affixes and develop a method to jointly model the role they play in allowing inferences about the inflected forms of Spanish verbs.

2.2 Previous Work

A major goal of this research is to unify two distinct camps of literature related to paradigm shape. The first is historical and formal work that models stem space organization (roughly, how stem alternants are distributed in a paradigm) on an abstract level using traditional methods of morphological analysis. The second encompasses work on inflectional complexity that utilizes information-theoretic measures to quantify the predictability of inflected forms. These two areas of literature differ both in their emphases and in their methodologies, but both relate to paradigm shape insofar

pattern during their development (Maiden, 2005, p.146) as they concern the distribution of implicative relations within a paradigm. Relatively little work has been done to integrate the two perspectives—a gap this research seeks to fill.[4] Following is an overview of the main insights I build on from both strands of literature, along with the limitations that I seek to improve upon.

2.2.1 Stem Space Organization

The literature on stem space organization focuses on modeling the distribution of stem alternants within a paradigm and has been approached from both historical and theoretic perspectives. A prime example of the historical approach is Maiden (2005)'s analysis of how Romance stem alternant distributions were reinforced and extended over time, as discussed above. An example of the theoretic approach is Boyé and Cabredo Hofherr (2006)'s identification of eleven stem "zones" for Spanish verbs. These zones partition the paradigm into sets of cells that always exhibit the same stem form, providing an abstract basis for analyzing patterns across this organizational space. Boyé and Cabredo Hofherr (2006) argue that the distribution of Spanish stem alternants is predictable; certain cells predictably have the same form due to the systemic constraints of the stem space organization. More broadly, their and other formal work shows that cells with different stem forms enter into predictable implicative relations, which often have parallel distributions across classes.

To make this more concrete, looking again at the data in Table 1, I can derive the implication that if the PRS.1SG has stem alternation *-ie-* then the PRS.1PL has stem alternation *-e-*. This holds true for all the verbs in Table 1. There is also a parallel implication that if the PRS.1SG has stem alternation *-ue-* then the PRS.1PL has stem alternation *-o-*. Expanding these principles to the Spanish verbal inflection system, this example highlights that predictable implications hold among cells with different stem forms and that these implications are distributionally parallel across classes even when their phonological particulars differ. These are key insights I build on to formalize the idea of paradigm shape.

However, one limitation to the stem space organization work is that it does not quantify the ex-

[4]The work of Stump and Finkel (2013, 2015) is a notable example of research that also bridges this gap. I discuss it further below.

tent to which there is similarity in the implicative relations holding among inflected forms. Maiden (2005)'s N-pattern, L-pattern, and U-pattern are completely discrete; a lexeme's paradigm either exhibits them or it does not. Since MOVER and SENTIR both have the N-pattern, they are considered the same with regard to this distribution. But consider that DECIR 'say' both shares the N-pattern and also follows the L-pattern.[5] In this case, DECIR is similar to MOVER and SENTIR in that they all have the N-pattern, but it is not precisely identical to them due to the additional presence of the L-pattern. Discrete categorization based on the three patterns does not encapsulate such insights about partial similarity. Likewise, Boyé and Cabredo Hofherr (2006)'s approach does not indicate the extent to which words have similar implicative relations; a lexeme either has a stem alternant in a particular zone or it does not. I seek to move beyond discrete categorization of stem alternations to gradient analysis that captures overlapping and partial similarities.

Another limitation is the focus on stem organization to the detriment of affix distributions. Maiden (2005)'s patterns and Boyé and Cabredo Hofherr (2006)'s zones both exclusively concern patterns in stem forms. Inflectional affixes also have distributions that impact paradigm shape, most notably in the theme vowels that are in many inflected forms. For example, in Table 1 SENTIR has affixes with the theme vowel *-e* in the 2SG, 3SG, and 3PL and affixes with the theme vowel *-i* in the 1PL and 2PL. This affix distribution has bearing on how predictable other forms of SENTIR are, as well as the extent to which SENTIR is similar to other microclasses. Given a form of SENTIR like *sientes* that has theme vowel *-e*, another form of SENTIR with theme vowel *-e* like *siente* is more predictable than a form with theme vowel *-i* like *sentimos*. Moreover, for the forms in Table 1, it is clear that SENTIR shares some affixal similarities with MOVER (which consistently has affixes with the theme vowel *-e* in the present indicative) but that the similarity is only partial because of SENTIR's distribution of theme vowels.

In addition, due to the formal morphological nature of stem space organization work, pre-analytical assumptions are often made about the inflectional system under analysis. For example,

[5]These patterns overlap in the 1SG indicative present, where the L-pattern takes precedence in the sense that the 1SG shares a stem alternant with the entire present subjunctive rather than with the 2SG, 3SG, and 3PL present indicative. I refer to verbs like DECIR that demonstrate this overlap as having a "modified L-pattern."

Boyé and Cabredo Hofherr (2006) draw a presumptive distinction between regular and irregular verbs and, within irregular verbs, they distinguish what they term "stem suppletions" (irregular but analyzable forms) and "form-suppletions" (irregular and unanalyzable). Such pre-analytical assumptions are helpful in developing categorical stem abstractions but can be counterproductive when trying to capture partial similarities in distributions that may hold across one or more of the pre-determined categories. Furthermore, the analytic assumptions are necessarily language-specific, so they also limit the extent to which the analysis can be replicated in new languages. Since this research seeks to license inferences about inflected forms in a way that is replicable in other languages, I jointly model distributional parallelism in stems and affixes without making any pre-analytic assumptions about the inflectional system at hand.

2.2.2 Inflectional Complexity

The inflectional complexity literature closely addresses predictability, often using conditional entropy to measure the uncertainty associated with predicting some unseen form. A key focus is on what Ackerman et al. (2009) termed the "Paradigm Cell Filling Problem": what licenses reliable inferences about the inflected surface forms of lexemes? For example, Ackerman and Malouf (2013, 2016)'s methods computed the complexity of a morphological system based on the average conditional entropy of a paradigm—the average amount of uncertainty associated with one random paradigm cell based on knowledge of another random cell. Under this analysis, the less reliably an inflectional system's networks of implicative relations enable inference of unobserved forms from observed forms, the more complex the system.

Since I am interested in quantifying the role that paradigm shape plays in the process of facilitating or inhibiting inferences about the inflected forms of lexemes, this type of work provides a computational starting point for answering this question. However, while average uncertainty like Ackerman and Malouf (2013, 2016) employed can provide a picture of a system's overall complexity, it obscures the distribution of uncertainty within an inflectional system, which is essential to evaluating paradigm shape. Furthermore, their approach abstracts away stem alternations to focus on affixal exponence. For example, under this approach, I might identify that some of the verbs in Table 1 have stem alternations and create abstract categories “stem” and “alt,” coding the PRS.2SG forms (in order) as stem+*as*, stem+*es*, alt+*as*, alt+*es*, and alt+*es*. Applied to the entire Spanish verbal system, such an abstraction would obscure not only the phonology of the alternations but also their precise distributions. As previously discussed, both stem and affix distributions are important to paradigm shape and need to be fully accounted for.

Stump and Finkel (2013, 2015) have also done extensive work on inflectional complexity, particularly in how principal parts structure inflectional systems. A principal part set is a minimal set of inflected forms that enables prediction of all other inflected forms of a lexeme. Viewed through the lens of implicative structure, principal parts are implicative relations that are highly informative about other inflected forms. Stump and Finkel (2015) delineate several different notions of principal parts, including static sets that require uniformity within a class and dynamic sets that allow variation across classes. This conceptualization enables investigation of distributional parallelism across lexemes and classes, since it captures the variation among paradigm cells in the extent to which they are informative about the exponence of other cells. However, similar to the stem space organization work, this approach encounters difficulty capturing partial parallelism due to the set-theoretic nature of principal part sets. This is evidenced by the fact that Stump and Finkel (2015) identify multiple “optimal” principal part sets for the inflectional systems they analyze.

A key insight for my work is Stump and Finkel (2015)’s concept of a “distillation,” a group of morphosyntactic property sets whose distinguishers (affix-like material) are isomorphic (i.e. have a one-to-one correspondence). In other words, this is an organization of classes based on exponent distribution rather than exponent phonology, which yields groupings of classes that have different exponents but in the same distribution. This is closely related to Baerman et al. (2017)’s distinction between “allomorphic” and “distributional” class systems. In an allomorphic system, classes are defined by inflectional exponents; if a given set of morphosyntactic properties is realized by different exponents in two lexemes, those lexemes belong to different classes. By contrast, in a distributional system, classes are defined by the distribution of exponents; if a morphosyntactic property set is realized by identical exponents in different distributions for two lexemes, they belong to different classes. Though these terms can be used definitionally to characterize inflectional systems, they are also useful in identifying different aspects of morphological structure within a single system. The Spanish verb system could be considered allomorphic because its classes are most commonly defined by inflectional exponents (specifically, theme vowels). At the same time, Maiden's N- and L-patterns in Romance highlight that Spanish stem classes are distributional in nature. Both types of structure are relevant and need to be captured for a full picture of paradigm shape in Spanish.

This poses a challenge to the information-theoretic methods that have been used in the inflectional complexity literature because conditional entropy can quantify allormophic inflection class systems much more readily than distributional ones. Entropy is calculated over surface exponents, but identifying distributional parallelism requires a degree of abstraction. The entropy approach is more suited to deduce, for example, that CANTAR and PENSAR in Table 1 are realized by the same affixal exponents in some cells than that PENSAR, MOVER, and SENTIR have parallel distributions of alternants realized by different phonological exponents. To address this challenge, my work develops a method of transforming the input data into a purely distributional form, a process referred to later on as "deidentification."

By building on work from literature on stem space organization and inflectional complexity, I apply information-theoretic methods to investigating distributional parallelism across lexemes and classes. I also improve upon the limitations of previous work by capturing partial similarity in paradigm organization, modeling the joint effects of stem alternations and affixes, and identifying groupings underlying both allormorphic and distributional class systems.

3 Method to Quantify Paradigm Shape

This section details the computational method I developed to quantify paradigm shape. Section 3.1 provides a conceptual overview of the method, section 3.2 explains the technical details of the algorithm, section 3.3 illustrates how the method works using a toy example, and section 3.4 covers the parallel method of "deidentification" developed to capture purely distributional information.

3.1 Method Overview

The method centers around a process of identifying "confusion sets" in order to quantify the strength of implicative relations between cells. These sets of cells "confuse" two microclasses in the system because internal comparisons based on the set do not allow precise assignment of a verb to one microclass or the other. To separate themes and distinguishers, local segmentation is performed in the style of Beniamine et al. (2017). The resulting distinguisher sets are used to identify confusion sets. Using entropy (Shannon, 1948), the method then computes the degree to which each confusion set helps identify the exact inflectional microclass of each verb. The resulting entropy values are structured in a matrix of m microclasses $\times$ n sets of cells, where each entry corresponds to the entropy value associated with a set of cells for a particular microclass. Implemented in Python, the method yields output that provides a quantitative basis for analyzing the inflectional system along multiple organizational dimensions.

3.1.1 Segmentation

The first step in executing the method is performing segmentation. Given a set of forms for a single lexeme, segmentation separates the *theme*, stem-like material that remains invariant for every form in the set, from the *distinguishers*, affix-like material that varies across the set. Despite the fact that segmentation-based analysis of morphological systems is common in the computational literature, there is no formal morphological consensus on how to perform a "correct" segmentation, even for very well-studied languages (Spencer, 2012). Careful choice of segmentation strategy is important because different strategies can lead to different analyses of inflectional structure. One choice in particular is pertinent to this work, the choice between a "global" segmentation strategy that identifies a single theme for each lexeme and a "local" segmentation strategy that identifies a theme for each set of a lexeme's forms. Beniamine et al. (2017) show that a local strategy produces better

LEXEME	PRS.1SG, PRS.2SG	THEME	DISTINGUISHERS
CANTAR	canto, cantas	cant	o, as
SUBIR	subo, subes	sub	o, es
SENTIR	siento, sientes	sient	o, es
PENSAR	pienso, piensas	piens	o, as
MOVER	muevo, mueves	muev	o, es

Table 3: Segmentation of PRS.1SG and PRS.2SG forms of verbs in Table 1

descriptions of lexemes with stem alternations; since this research is interested in stem alternations in addition to affixes, I follow their approach in using local segmentation.

Table 3 illustrates the segmentation process for the PRS.1SG and PRS.2SG forms of the verbs shown previously in Table 1. Taking CANTAR as an example, segmentation assigns the material shared by both forms (*cant*) to the theme and the material that differs between the two forms to the distinguisher set (*-o, -as*). The other four verbs are segmented similarly. It is important to note that, even though SENTIR, PENSAR, and MOVER have stem alternations in the present indicative, the set {PRS.1SG, PRS.2SG} only includes one stem allomorph for each of these verbs. Consequently, the stem allomorph is retained in the theme, meaning that analysis of this set can show regularity of affixes. For example, CANTAR and PENSAR have identical distinguishers for the set {PRS.1SG, PRS.2SG}, showing regularity in their affixes.

By contrast, sets that do contain multiple stem allomorphs highlight the presence of stem alternations. Table 4 parallels Table 3 but segments the PRS.1SG and PRS.1PL forms of the verbs.[6] Since the PRS.1SG contains one stem allomorph for SENTIR, PENSAR, and MOVER and the PRS.1PL contains another stem allomorph, the alternating stem characters are segmented into the distinguisher sets for these verbs. Thus, for the set {PRS.1SG, PRS.1PL}, CANTAR and PENSAR no longer have identical distinguisher sets due to the presence of the stem alternation in PENSAR. This illustration highlights that using local segmentation to analyze different sets of forms yields insights both about affixes and about stem alternations, a key aim of the method.

[6]For the sake of simplicity, these illustrations only involve sets containing two forms, but the segmentation process outlined here applies to sets of any size.

LEXEME	PRS.1SG, PRS.1PL	THEME	DISTINGUISHERS
CANTAR	canto, cantamos	canto	–, ams
SUBIR	subo, subimos	subo	–, ims
SENTIR	siento, sentimos	sento	i, ims
PENSAR	pienso, pensamos	penso	i, ams
MOVER	muevo, movemos	mvo	ue, oems

Table 4: Segmentation of PRS.1SG and PRS.1PL forms of verbs in Table 1

3.1.2 Maximal Confusion Sets

After segmentation, the next step is identifying confusion sets. A set of cells "confuses" two microclasses if local segmentation of the inflected forms for both classes yields identical distinguisher sets. Returning to Table 3, local segmentation of the PRS.1SG and PRS.2SG forms yields the same distinguisher set {*-o, -as*} for CANTAR and PENSAR. Consequently, the set {PRS.1SG, PRS.2SG} is a confusion set for these microclasses. This means that, given some Spanish verb, knowledge of its PRS.1SG and PRS.2SG forms does not allow precise identification of that verb to either the CANTAR class or the PENSAR class. By contrast, the set {PRS.1SG, PRS.1PL} from Table 4 does not yield the same distinguisher set for CANTAR and PENSAR, so it does not constitute a confusion set for these classes.

This notion of confusion enables the method to capture local similarities between microclasses that are by definition globally different. The utility of this approach is highlighted by its application even when a verb's stem is entirely suppletive. For example, SER 'be' has a suppletive preterite *fui*. Because identifying confusion sets relies on internal contrasts between sets of cells, comparing SER to another verb via a set that includes both a present form and a preterite form would capture the fact that SER's suppletive preterite is unique. However, comparison via a set that contained only preterite forms would highlight that SER has a relatively regular conjugation of preterite forms taken on their own.

Identifying every possible confusion set would mean enumerating every set of cells that confuses two microclasses. This is made difficult by the fact that, for every large set S that confuses two microclasses, all of its exponentially many subsets also confuse those two microclasses. For this reason, I restrict the method to identifying *maximal* confusion sets for each pair of microclasses.[7] A maximal confusion set is a set of cells *S* (size>1) which confuses microclasses *A* and *B* and for which no superset of *S* also confuses *A* and *B*. Further explanation of how the method efficiently computes maximal confusion sets is included in section 3.2.

3.1.3 Entropy

Once all maximal confusion sets have been identified for the inflectional system under examination, the next step is evaluating each set's predictive power using entropy. Entropy is a measure of uncertainty drawn from information theory (Shannon, 1948) that this method uses to compute how well a maximal confusion set enables precise identification of each microclass's identity. If a given microclass is confusable with *k* microclasses on the basis of some set of cells, the remaining uncertainty about that microclass's identity can be quantified as $\log_2 k$ bits.[8]

For example, in Table 3, the set of cells {PRS.1SG, PRS.2SG} partitions the given five microclasses into two mutually confusable partitions. Since CANTAR and PENSAR both have the distinguisher set {*-o, -as*}, each of them can be confused with two microclasses. This means that their count *k* equals 2 and the remaining uncertainty for both of them is $\log_2 2 = 1$ bit. Since SUBIR, SENTIR, and MOVER all have the distinguisher set {*-o, -es*}, each of them can be confused with three microclasses. This means that their count *k* equals 3 and the remaining uncertainty for each of them is $\log_2 3 = 1.585$ bits. In the case that a set of cells yields a unique distinguisher set for a particular class, that class's count *k* would equal 1 and its entropy would be $\log_2 1 = 0$ bits. This final case highlights entropy's usefulness as a quantitative standard; when a set of cells uniquely identifies a particular microclass, that class is assigned an entropy value of zero, reflecting that no uncertainty remains about its identity.

All steps of the method were implemented in Python. Entropy is applied to each maximal confusion set and each microclass. The output matrix is structured such that each entry corresponds

[7] Computing maximal confusion sets for more than two microclasses at a time is left to future work.

[8] Note that this quantification implicitly assumes that all the microclasses are equally likely. In practice, inflection classes in a given system do not tend to all be equally likely.

to the entropy value associated with a set of cells for a particular microclass. This matrix of entropy values quantifies the distribution of predictive relationships across the inflectional system. Further details of the algorithm are included below.

3.2 Details of Algorithm

This section steps through the algorithm used to implement the original method, explaining how it efficiently computes maximal confusion sets for each pair of microclasses.

Conceptually, the process starts with the insight that every maximal confusion set must be associated with a theme for each microclass it confuses. If a given morphosyntactic property set can yield an identical distinguisher set for class I and class J, then there must be some theme A for class I and some theme B for class J that is segmented away to produce these distinguishers. For a more concrete example, consider that the CANTAR and PENSAR classes in Table 3 produce the same distinguisher set $\{$*-o, -as*$\}$. The fact that they do so presupposes the existence of themes that produce these distinguishers under local segmentation. In this case, we know that the corresponding themes are *cant* and *piens*, but the logic holds even when the accompanying themes are unknown.

Following from this reasoning, the first step in the algorithm is identifying all possible themes for every microclass in the data. Since themes are the longest possible character subsequences that occur in every form in a set, they grow monotonically shorter as more inflected forms are added. Since the form set $\{$canto, cantas$\}$ yields theme *cant*, the addition of another form to the set can result in a theme that is equal to or shorter than *cant* in length, but not longer. This means that all possible themes for a given microclass can be computed by first aligning every pair of forms within the class and then aligning the resulting themes until no additional themes can be generated. This process of identifying all possible themes for a class is followed for every microclass in the data.

Given all the possible themes for a class, the algorithm identifies the forms in the class for which each theme is valid. This is done by comparing each theme against each form in the class and finding the largest possible set of forms for which the theme is valid. This set can be denoted as $S(A)$ for some theme A. After this has been done for every microclass in the data, the next step is to identify maximal confusion sets. Given the largest possible form sets for all the themes in class I and the largest possible form sets for all the themes in class J, the pairwise intersections of these form sets will necessarily yield all maximal confusion sets (along with some non-maximal confusion sets). That is, pairing every form set identified for class I with every form set identified for class J is guaranteed to capture every set of forms that confuses class I and class J and for which no superset also confuses class I and class J.

To find confusion sets for classes I and J, the form sets for every pair of themes A_i and B_j are tested. The method takes the set $S(A_i) \cap S(B_j)$ and tests whether it has at least two members; if so, local segmentation is performed to check that the set members actually produce the themes A_i and B_j. This check is important because the intersection may be smaller than the original sets and local segmentation could therefore produce themes that are longer than A_i and B_j, which would make the set invalid for A_i and B_j. All sets $S(A_i) \cap S(B_j)$ that have $\geq$2 members and pass the check are output as potential maximal confusion sets. After all such potential sets are generated for a pair of classes, any sets that are subsets of other sets are removed to ensure that only maximal confusion sets are retained. This process of identifying maximal confusion sets is followed for every pair of classes in the system. The results are structured in a matrix of m microclasses $\times$ n maximal confusion sets.

After this matrix has been generated, the final step of the algorithm is applying entropy. For each maximal set in the matrix, the method iterates through the microclasses, calculating for each class how many classes it can be confused with based on the forms in the maximal set. Two classes are confusable by a maximal set if it is possible for both classes to generate an identical distinguisher set for the forms in that maximal set. For each given maximal set, the count *k* for every class starts at 1 (since a class can always be confused with itself) and increases by 1 for each additional class it can be confused with based on that maximal set. The corresponding cells in the matrix are filled with the entropy of these count values, $\log_2 k$. This yields the final result of the method, a matrix of m microclasses $\times$ n maximal confusion sets where each entry represents the uncertainty that remains about a particular microclass given some maximal set.

3.3 Toy Example

This section illustrates how the method works on a small data set. Table 5 shows the aorist, imperfective, and perfective forms of nine verbs in Gulmancema, a language in the Niger-Congo family spoken in the easternmost parts of Burkina Faso (ISO 639-3 code `gux`). According to Baerman et al. (2017), the affix-mood suffixes of verbs in Gulmancema represent a union of both allomorphic and distributional classes. The paradigm of each verb involves the presence or absence of one of the suffixes *-ni*, *-di*, *-li*, and *-gi*. This constitutes the allomorphic aspect of the class structure, since morphosyntactic property sets can be realized by different exponents in different lexemes. However, the suffixes can occur in any of three distributions (the imperfective, the perfective, or the aorist and perfective), and some verbs are entirely unsuffixed. This makes the class structure also distributional, since morphosyntactic property sets can be realized by identical exponents in different distributions in different lexemes. This combination of allomorphic and distributional class structure makes the Gulmancema data a useful toy example on which to test my method of quantifying paradigm shape.

Table 6 shows the results of the method for the Gulmancema data. The method identified three maximal confusion sets for the data: {AOR, PFV}, {IPFV, PFV}, and {AOR, IPFV}. For each of these maximal confusion sets, each verb has an associated entropy value. As an example, for the {IPFV, PFV} maximal confusion set, TUA 'tap on the head' has an entropy value of 0. This means that the set {IPFV, PFV} precisely identifies the class of this verb. Looking at the data, it is clear that TUA is the only verb that has the *-ni* suffix in the imperfective and no suffix in the perfective. Therefore, the count of classes it could belong to is 1 and the remaining uncertainty about its identity is 0. By contrast, for the same maximal confusion set, GOA 'return' has an entropy value of 1, meaning that there are two possible options for the class it could belong to. This verb is unsuffixed in the imperfective and has the suffix *-ni* in the perfective. From the data, it is clear that the verb TIE 'do' also has this distributional pattern. Thus, the count of classes that the verb could

	'tap on the head'	'return'	'do'	'pass'	'love'	'hear'	'fall'	'give birth'	'plant'
AOR	tua	goa	tie-ni	cie	bua	gba-di	ba	ma	bu-li
IPFV	tua-ni	goa	tie	cie-di	bua	gba	baa-li	ma	bu
PFV	tua	goa-ni	tie-ni	cie	bua-di	gba-di	ba	ma-li	bu-li

Table 5: Gulmancema data from Baerman et al. (2017, p. 116), originally from Naba (1994) and Ouoba (1982); tonal markings are removed for simplicity

	AOR, PFV	IPFV, PFV	AOR, IPFV
'tap on the head'	2.5849625	0	0
'return'	0	1	1.5849625
'do'	2.5849625	1	0
'pass'	2.5849625	0	0
'love'	0	1	1.5849625
'hear'	2.5849625	1	0
'fall'	2.5849625	0	0
'give birth'	0	1	1.5849625
'plant'	2.5849625	1	0

Table 6: Matrix generated by applying method to Gulmancema data

belong to is 2 and the remaining uncertainty about its identity given the set {IPFV, PFV} is 1.

All of the other entropy values in Table 6 are arrived at in the same manner. If a verb is precisely identified by a maximal confusion set, its entropy is 0. If not, its entropy value reflects the number of other classes it could be confused with. Taken as a whole, the results provide a gradient, numerical characterization of the structural similarities between the verbs in this toy data set. However, it is also important to note what information this method is not capturing. The method can identify that the form sets {*goa*, *goa-ni*} and {*tie*, *tie-ni*} have the same distribution because they share the same suffix *-ni*. For form sets that have the same distribution but different suffixes, like {*goa*, *goa-ni*} and {*bua*, *bua-di*}, the method treats them as formally independent. In other words, the method can identify only some distributional class structure. In order to capture distributional parallelism of forms realized by different phonological exponents, abstracting away from the surface forms of those exponents is required. This task is addressed in the next section via the development of a "deidentified" method that parallels the original one described until this point.

The Gulmancema toy example also highlights methodological limitations related to how realistic the method's assumptions are with respect to speakers' knowledge and inferences. For example, the set {AOR, PFV} confuses six of the Gulmancema verbs. Three of these verbs are suffixed in both the aorist and the perfective, like TIE, which has *tie-ni* for both tenses. The remaining three are suffixed in neither tense, like TUA, which has *tua* in both the aorist and the perfective. Since local segmentation of two identical forms yields an empty distinguisher set, the method treats both of these situations—suffixed in both tenses and suffixed in neither—identically. While this does accurately identify that all six verbs have identical aorist and perfective forms, it seems unlikely that speakers would conceptualize a suffix like *-ni* as belonging to a verb's theme, even if the suffix occurred in every form they knew. Though some distributional parallels across paradigms are certainly cognitively real to speakers, as previously discussed, not all are. For this reason, the similarities in paradigm shape across inflection classes which my model captures should not automatically be taken as reflective of speaker knowledge without further evidence to that end.

3.4 Deidentification

The method described thus far detects differences between microclasses based on the affixes in a set and the distribution of stem allomorphs within a set. However, there still remains additional distributional information that is not being captured. For example, the distinguisher sets {*-o, -as*} and {*-o, -es*} for CANTAR and MOVER, respectively, in Table 3 are not identical in phonological exponents but they are parallel in the distribution of those exponents. Since the method compares the surface forms of the distinguisher sets, it is unable to capture parallel distributions realized by different phonological exponents.

To account for this, I also develop a second method of "deidentification" that abstracts away from the forms of the distinguishers to focus on purely distributional information. To achieve this, the individual characters within the distinguisher sets are replaced with abstract identifiers indicating the position of identical characters. For example, the distinguisher set {*-o, -as*} would be represented as $\{\alpha, \beta\gamma\}$; since {*-o, -es*} would be represented the same way, the two sets can be matched to highlight that the distribution of the affixal exponents is identical despite the fact that they have different theme vowels.

In order to replace the distinguisher characters with abstract identifiers, the sets of forms compared to generate confusion sets must be "deidentified." The deidentified method is exactly the same as the original method up until this point—all possible themes are identified for each class and the largest possible form sets are determined for each theme. Before form sets from two different classes are compared, a multi-string alignment is performed on the distinguishers of both sets to replace their characters with abstract identifiers, as previously described. I used the A^* algorithm to search for multi-way alignments between the distinguishers as an efficient means of searching for an optimal multi-string alignment (Russell and Norvig, 2021). It is important to note here that local segmentation can produce multiple distinguishers for a given form. For any form in the form sets with multiple possible distinguisher options, the method randomly selects one of these options to be added to the distinguisher set and have the alignment performed with. This makes the multi-way alignment an approximation of the ideal multi-way alignment between the two form sets, with the potential for different alignments to be generated in successive runs on the same form sets.

After converting the distinguishers of both form sets into abstract identifiers, the next step is identifying confusion sets between microclasses. The original definition of confusion laid out in section 3.1.2, which was based on distinguisher sets being identical, requires modification for the deidentified case. For the deidentified analysis, a set of cells "confuses" two microclasses if local segmentation of the inflected forms for both classes yields deidentified distinguishers for which a perfect one-to-one correspondence is possible. This modified definition enables identification of deidentified sets that are distributionally identical but have been assigned different abstract identifiers. For example, the deidentified distinguisher sets $\{\alpha, \beta\gamma\}$ and $\{\gamma, \delta\epsilon\}$ clearly have the same distribution even though they are not identical, and the deidentified method is able to identify the perfect one-to-one correspondence between them (α:γ, β:δ, γ:ϵ). To identify one-to-one correspondences that maximize the number of deidentified forms that match between a pair of classes, this method once again uses the A^* algorithm.

After all the potential confusion sets are generated for a pair of classes, any sets that are subsets of other sets are removed, keeping only the maximal confusion sets. These maximal confusion sets are structured in a matrix as before. Given a maximal confusion set and a particular class, calculating how many classes that class can be confused with based on the forms in the maximal set relies on the modified definition of confusion. Two classes are confusable by a maximal set if it is possible to establish a perfect one-to-one correspondence between both classes' deidentified distinguisher sets for the forms in that maximal set. After computing the count *k* for each maximal set for each class, entropy is applied in the same manner as the original method, yielding a final matrix of m microclasses $\times$ n deidentified maximal confusion sets.

4 Results with Spanish Verbs

For my thesis, I applied both the original and the deidentified methods to the Spanish verbal inflectional system, using 60 morphosyntactic property sets of 40 Spanish verb microclasses (drawn from Brodsky, 2005). The original method generated 290 maximal confusion sets for this data, for a resulting 40 $\times$ 290 matrix of entropy values. The deidentified method generated 25,239 maximal confusion sets, for a resulting 40 $\times$ 25,239 matrix of entropy values.

These matrices of entropy values were visualized using t-SNE, a technique that enables visualization of high-dimensional datasets (van der Maaten and Hinton, 2008). With t-SNE, each matrix is visualized as a plot of 40 points, one for each microclass. Figure 1 shows the t-SNE visualization for the maximal confusion sets generated by the original method, and Figure 2 shows the same for the deidentified method. Both plots are coded by color according to the traditional Spanish verb class categorizations and by shape according to Maiden (2005)'s alternation patterns. To determine the traditional class categorizations, I grouped all the classes Brodsky (2005) deemed "fundamentally irregular" together and then organized the remaining "basically regular" classes into *-ar*, *-er*, and *-ir* groups based on their infinitive forms. For the alternations, I identified the L-pattern, the N-pattern, and a "modified L-pattern" (a mixed N-pattern and L-pattern) in the data.[9]

[9]In addition to t-SNE, I also used hierarchical clustering to create visualizations of the two matrices. These are

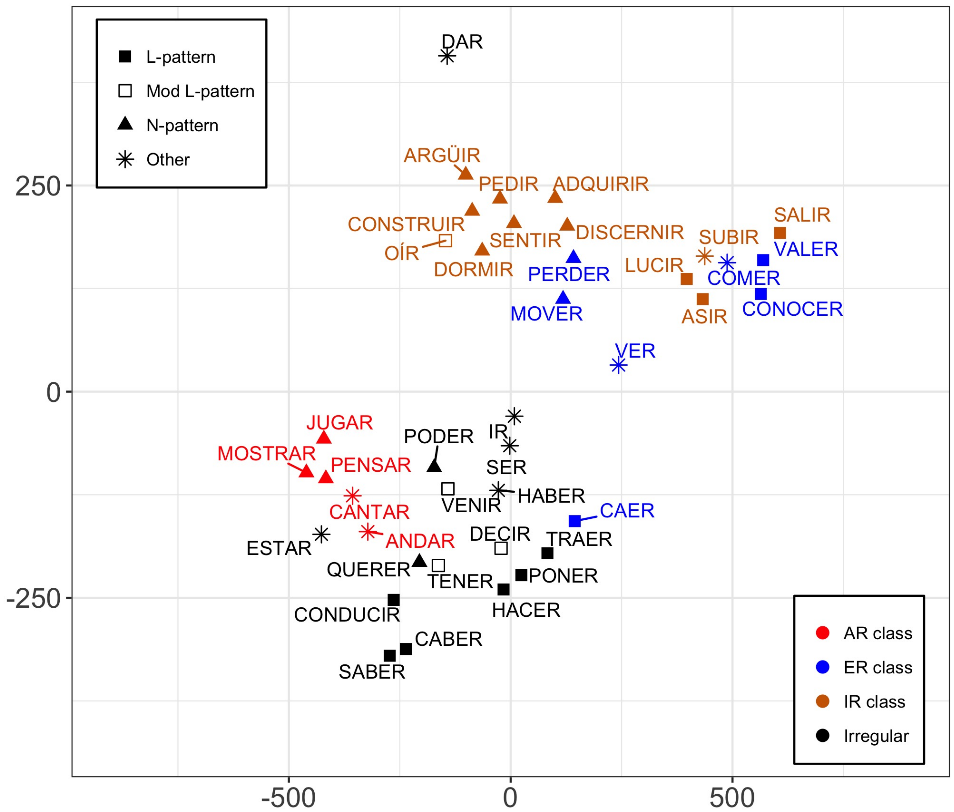

Figure 1: Results of t-SNE analysis based on entropy of maximally confusable sets. Colors show the traditional classes; symbols show the Maiden alternation patterns.

These visualizations highlight several key components of paradigm shape in Spanish verbs. Traditional allomorphic class groupings are readily distinguishable in Figure 1, most clearly in the cluster of red *-ar* verbs. By contrast, the *-er* and *-ir* verbs are somewhat interspersed. This comports with the fact that the *-ar* classes retain the theme vowel *a* in their suffixes fairly consistently across their paradigms, while the *-er* and *-ir* classes demonstrate inconsistency in the realization of an *i* vs. *e* suffix (cf. the discussion of SENTIR's affix distribution in section 2.2.1). These clustering structures show that the method is successfully capturing aspects of allomorphic classes defined by inflectional exponents—in particular, theme vowels.

Distributional class groupings are also present. For example, the large swath of classes at the

presented in the form of dendrograms, in which the points are clustered using the complete method from Scikit Learn (Pedregosa et al., 2011). These highlight the same underlying clusters I have described in the t-SNE plots and are included in Appendix A.

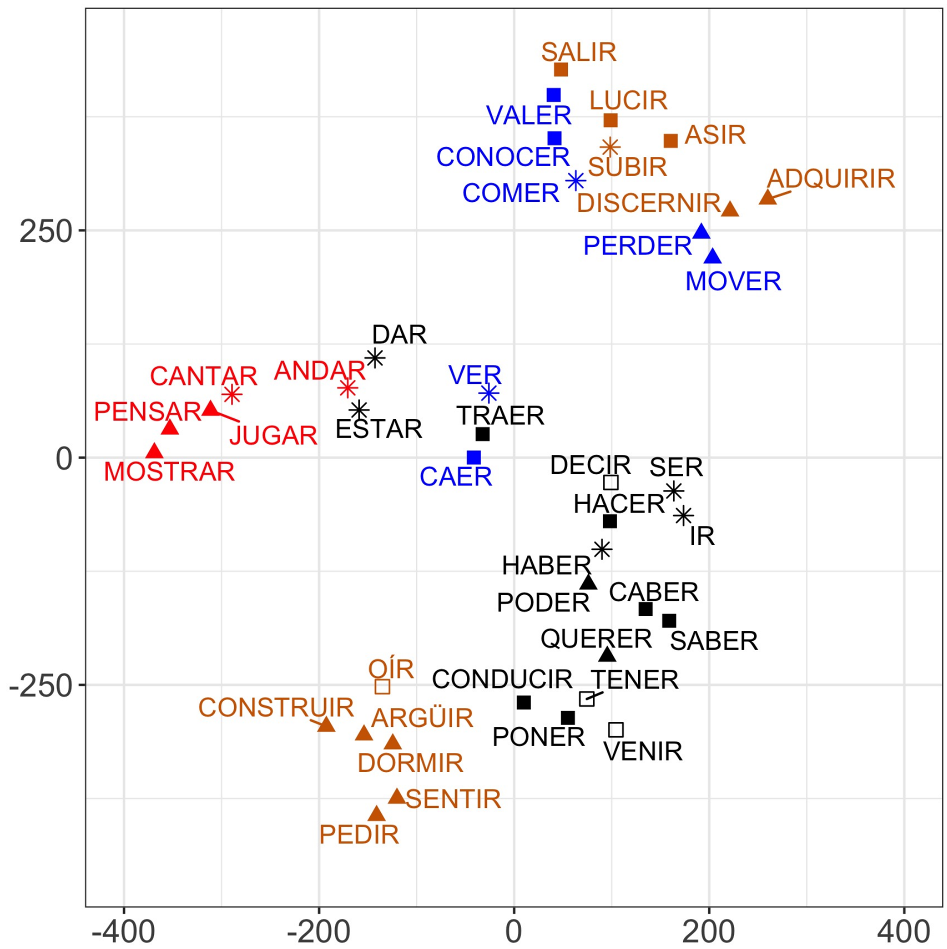

Figure 2: Results of t-SNE analysis based on entropy of maximally confusable sets in the deidentified condition. Coding of colors and shapes is the same as in Figure 1.

top of Figure 1 has two main clusters. Each consists of both *-er* and *-ir* verbs, so it is clear that the clustering is not driven by inflectional affixes alone. In fact, Maiden (2005)'s stem alternations explain most of this clustering distinction. Most verbs in the left-hand cluster exhibit the N-pattern (SENTIR 'feel', PEDIR 'ask for', DORMIR 'sleep', CONSTRUIR 'build', ARGÜIR 'argue', OÍR 'hear', PERDER 'lose', MOVER 'move', DISCERNIR 'discern', and ADQUIRIR 'acquire') while most in the right-hand cluster have the L-pattern (LUCIR 'shine', ASIR 'grab', CONOCER 'know', COMER 'eat', SUBIR 'rise', VALER 'cost', and SALIR 'leave'). Though the distinction is not perfect, it clearly shows that the method is identifying some distributional distinctions.

My deidentified approach is able to take this one step further and draw even finer distributional

LEXEME	GLOSS	PRET.1SG	PRET.2SG	PRET.3SG	PRET.1PL	PRET.2PL	PRET.3PL
SENTIR	'feel'	sentí	sentiste	sintió	sentimos	sentisteis	sintieron
PEDIR	'ask for'	pedí	pediste	pidió	pedimos	pedisteis	pidieron
DORMIR	'sleep'	dormí	dormiste	durmió	dormimos	dormistes	durmieron
CONSTRUIR	'build'	construí	construiste	construyó	construimos	construisteis	construyeron
ARGÜIR	'argue'	argüí	argüiste	arguyó	argüimos	argüisteis	arguyeron
OÍR	'hear'	oí	oíste	oyó	oímos	oísteis	oyeron
PERDER	'lose'	perdí	perdiste	perdió	perdimos	perdisteis	perdieron
MOVER	'move'	moví	moviste	movió	movimos	movistes	movieron
DISCERNIR	'discern'	discerní	discerniste	discernió	discernimos	discernisteis	discernieron
ADQUIRIR	'acquire'	adquirí	adquiriste	adquirió	adquirimos	adquiristeis	adquirieron

Table 7: Preterite alternation that leads to the cluster split observable in Figure 2

distinctions between classes. The previously mentioned upper-left-hand cluster in Figure 1 splits into two smaller clusters under the deidentified method in Figure 2. Though the large cluster is united by all of its members having Maiden's N-pattern (except for OÍR, which has the mixed N-pattern and L-pattern), the split into smaller clusters can be explained by another alternation in the preterite. As shown in Table 7, the verbs in the first group (SENTIR, PEDIR, DORMIR, CONSTRUIR, ARGÜIR, and OÍR) all have an alternation in their third person singular and third person plural preterite indicative forms, whereas those in the second group (PERDER, MOVER, DISCERNIR, and ADQUIRIR) have no alternations in the preterite. This difference seems to motivate the distributional split.

Significantly, the verbs in the first group do not all exhibit the same alternation: SENTIR and PEDIR have *e*∼*i*; DORMIR has *o*∼*u*; and CONSTRUIR, ARGÜIR, and OÍR have *i*∼*y*. This highlights one of the benefits of the local segmentation strategy. While the *e*∼*i* and *o*∼*u* alternations occur straightforwardly in the stem, the *i*∼*y* alternation appears at the boundary between stem and affix—that is, it could plausibly be segmented into either the stem or the affixes. Since local segmentation treats variant material the same regardless of whether it is part of a stem or an affix, the *i*∼*y* can be grouped with the other two alternations that occur in comparable positions. This illustrates the strength of the deidentified method is identifying purely distributional class structure, uncovering structural similarities despite different phonological exponents.

Overall, the results show that the method developed in this thesis provides a quantitative basis for identifying structural similarities between the paradigms of Spanish verbs. In doing so, it captures several pre-existing intuitions about the implicative structure of Spanish verbal inflections, including both the traditional inflection classes and Maiden (2005)'s distributional classes. Moreover, it also makes finer distinctions which were not explicitly listed in prior work, but which follow from their principles of analysis.

5 Conclusion

This thesis presents a method for precisely quantifying paradigm shape in a replicable, extendable way. Bridging the gap between formal literature on stem space organization and information-theoretic literature on inflectional complexity, this work models partial parallelism in the implicative relations holding among inflected forms by accounting for the effects of both stem alternations and suffixes. By identifying maximal confusion sets for pairs of classes, the computational method developed in this research is able to use entropy to quantify structural similarities between classes. In addition, the parallel deidentified method abstracts away from surface forms to focus on purely distributional information. Applied to Spanish verbs, the original method successfully captures both allomorphic and distributional class insights, and the deidentified method draws even more fine-grained distributional distinctions between classes.

In the future, this work could be improved and expanded upon in several ways. First, the conceptualization of confusion sets delineated here only accounts for two microclasses at a time. It would be useful to develop a notion of confusion sets that could compare multiple microclasses, as this would provide an even fuller picture of paradigm shape than laid out here. In addition, as noted, my method's entropy calculations assume that all the inflection classes are equally likely, an assumption that is generally untrue. In most inflectional systems, some classes have more members than others and are therefore more likely. This issue could be addressed by incorporating type frequency information into the method's calculations (Sims and Parker, 2016; Parker and Sims, 2019).

Furthermore, there are several morphological phenomena my method is ill-equipped to handle. Since segmentation is performed early in the process and class comparison is based only on distinguisher sets, no information about the theme contributes to the method's evaluation of paradigm shape similarities. This means that the method will miss important information in cases where the phonological shape of a verb's theme is informative about its distinguishers. Another segmentation challenge is posed by systems with extensive reduplication, since they can yield many-to-one or many-to-many string alignments (Dolatian and Heinz, 2020). Particularly given the approximations the model uses to achieve multi-way alignments, it is unlikely to produce useful results when applied to cases of reduplication. Finally, since it was designed with Spanish in mind, the method is most readily extendable to other highly inflected languages and it is unclear how effective it would be for highly agglutinative ones. A fruitful line of future work would be applying the method to known problematic systems, like Tagalog verbs and Nuer nouns (Baerman, 2012; Baerman and Monich, forthcoming), to shed light on the precise limits of its usefulness when extended to non-Romance languages.

Ultimately, the computational method developed in this research was designed to be extended. Running it on additional inflectional systems would both test its utility with other languages and provide a quantitative basis for comparing paradigm shapes across languages. From a historical standpoint, it would also be of particular interest to run the method on Latin verbs and perhaps one or two other modern Romance verbal systems. This could lend insight into how paradigm shape has developed over time in the Romance family and impacted the historical development of Romance verbs.

Acknowledgements

I am greatly indebted to my advisors, Andrea Sims and Micha Elsner, for their ongoing mentorship and support throughout this research process, all the way from the initial development of the research question to the final project. Their guidance, kind encouragement, and constructive criticism have been invaluable in shaping my development as a researcher and decision to pursue graduate study in linguistics. I am truly grateful.

I would also like to thank Bob Levine and the members of the Autumn 2019 undergraduate research seminar here at Ohio State for their feedback on a preliminary version of this work. Additional thanks are owed to three anonymous SCiL reviewers for their comments on a conference paper based on this research.

References

Farrell Ackerman and Robert Malouf. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464, 2013.

Farrell Ackerman and Robert Malouf. Implicative relations in word-based morphological systems. *Cambridge handbook of morphology*, 297:328, 2016.

Farrell Ackerman, James P. Blevins, and Robert Malouf. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. *Analogy in grammar: Form and acquisition*, pages 54–82, 2009.

Matthew Baerman. Paradigmatic chaos in nuer. *Language*, pages 467–494, 2012.

Matthew Baerman and Irina Monich. Paradigmatic saturation in nuer. *Language*, forthcoming.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. *Morphological complexity*. Cambridge University Press, 2017.

Sacha Beniamine, Olivier Bonami, and Benoît Sagôt. Inferring inflection classes with description length. *Journal of Language Modelling*, 5:465–525, 2017.

Olivier Bonami and Sacha Beniamine. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182, 2016.

Gilles Boyé and Patricia Cabredo Hofherr. The structure of allomorphy in Spanish verbal inflection. *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset*, 13:9–24, 2006.

David Brodsky. *Spanish verbs made simple(r)*. University of Texas Press, 2005.

John Butt, Carmen Benjamin, and Antonia Moreira Rodríguez. *A new reference grammar of modern Spanish, 6th edition*. Routledge, 2019.

Hossep Dolatian and Jeffrey Heinz. Computing and classifying reduplication with 2-way finite-state transducers. *Journal of Language Modelling*, 8(1):179–250, 2020.

Martin Maiden. Morphological autonomy and diachrony. In Geert Booij and Jaap van Marle, editors, *Yearbook of morphology 2004*, pages 137–175. Springer, 2005.

Jean-Claude Naba. *Le Gulmancema: essai de systématisation; phonologie-tonologie-morphophonologie nominale-système verbal*. PhD thesis, Köppe Köln, 1994.

B Ouoba. *Description systématique du Gulmancema*. PhD thesis, Institut National des Langues et Civilisations Orientales, 1982.

Jeff Parker and Andrea D Sims. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of russian nouns. *InP. Arkadiev & F. Gardani Eds. The Complexities of Morphology*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Stuart Russell and Peter Norvig. *Artificial intelligence: A modern approach*. Pearson, 2021.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Andrea D Sims and Jeff Parker. How inflection class systems work: On the informativity of implicative structure. *Word Structure*, 9(2):215–239, 2016.

Andrew Spencer. Identifying stems. *Word Structure*, 5:88–108, 2012.

Gregory Stump and Raphael A Finkel. The complexity of inflectional systems. *Linguistics Vanguard*, 1(1):101–117, 2015.

Gregory T. Stump and Raphael A. Finkel. *Morphological typology: From word to paradigm*. Cambridge University Press, 2013.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Wolfgang U. Wurzel. *Inflectional morphology and naturalness*. Kluwer, 1989.

Appendix A: Dendrograms

The points in these dendrograms were clustered using the complete method from Scikit Learn (Pedregosa et al., 2011) and show the same basic structures discussed in the Results section. Figure 3 shows the results using the original method and Figure 4 shows the results using the deidentified method.

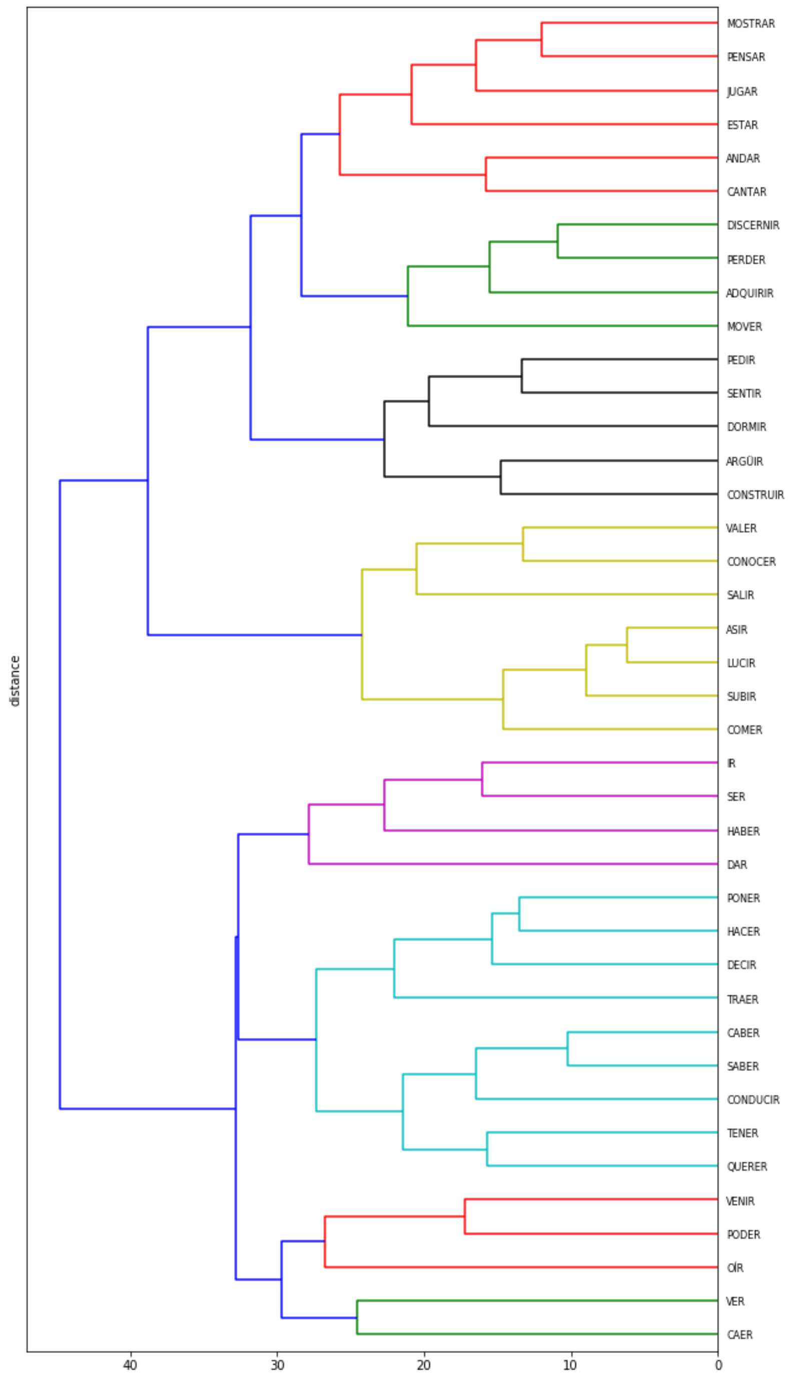

Figure 3: Results of hierarchical clustering analysis based on entropy of maximal confusion sets under the original method.

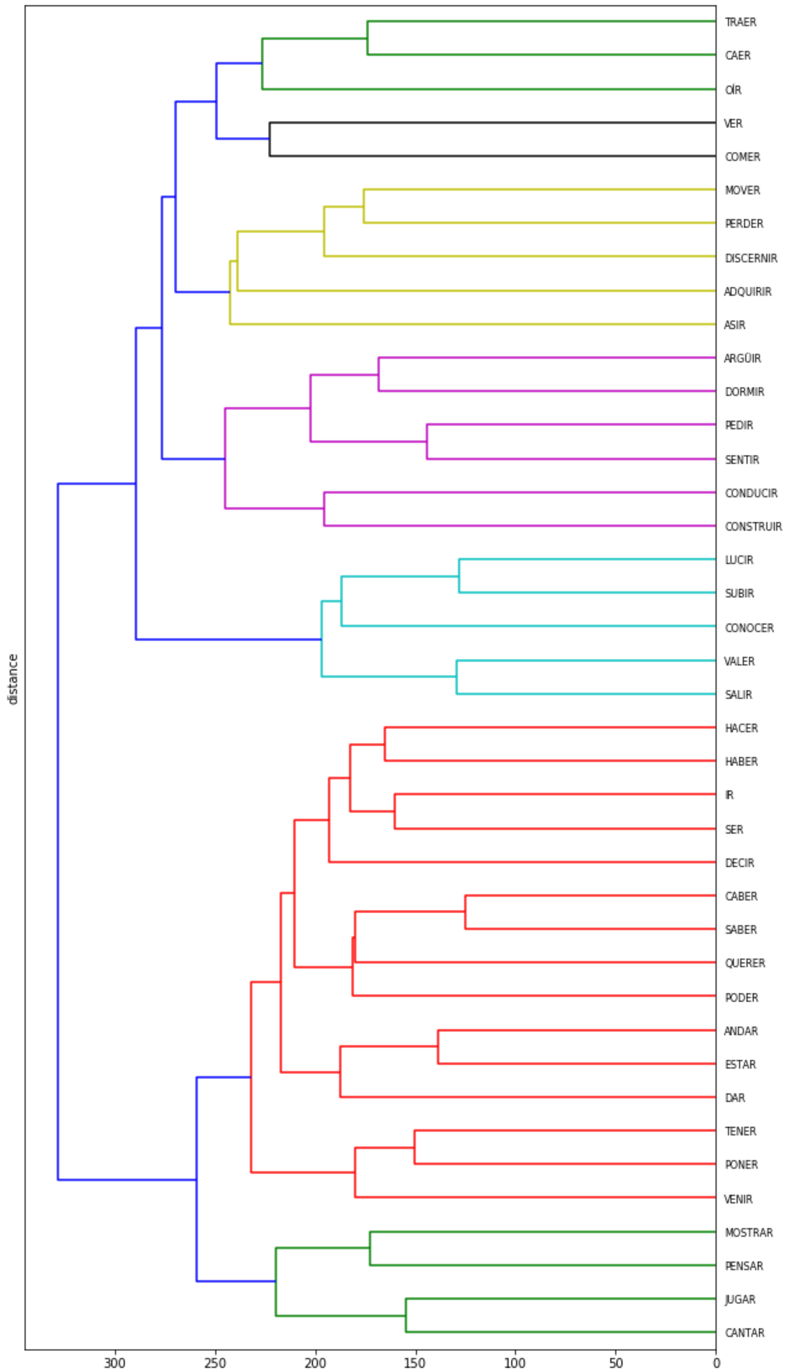

Figure 4: Results of hierarchical clustering analysis based on entropy of maximal confusion sets under the deidentified method.